ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 10, Issue, 05(E), pp. 32494-32497, May, 2019

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

FIGHT AGAINST PHISHING: A DATA MINING TECHNIQUE

Rahul Patel and Ananad Rajavat

Shri Vaishnav Institute of Information Technology SVVV Indore

DOI: http://dx.doi.org/10.24327/ijrsr.2019.1005.3483

ARTICLE INFO

Article History:
Received 13th February, 2019
Received in revised form 11th March, 2019
Accepted 8th April, 2019
Published online 28th May, 2019

Key Words:

phishing attack, malicious URL classification, association rule mining, rule based classification, ARM algorithm, outlier removal.

ABSTRACT

The internet and applications are growing in a rapid rate. Additionally a number of new users are involving to consume services offered by the internet. On the other hand the cases of online fraud and phishing are also growing continuously. In this paper the cyber security and phishing is the key area of investigation and system design, more specifically the phishing attack. The phishing attack is deployed to target an innocent user. By which the target user can lose their financial status or social credibility. The phishing attacker still the user's private, sensitive and confidential data and using this attacker harm the target person. Therefore this paper is providing the study about phishing attack additionally the recently developed techniques that are useful for detection and prevention of the phishing attack. Finally we propose a data mining based technique that is used to identify the phishing attack.

INTRODUCTION

Internet and it applications become routine work now in these days. Additionally new applications and services are enabled for providing the services 24X7. All these things become possible by using the communication and technology advancement. On average a common person interacted with the internet many times in a day. But in the similar manner the cases of online fraud and cyber crime rate is also increases. Among the phishing is a one of most frequent attack in online fraud cases. Basically it is a cyber crime, in which innocent user passes own confidential and private data to attacker. Attacker misuse that gained information to harm the data owner. Therefore it is not only cyber crime it is also a serious social crime [1].

The phishing attack not only creates financial issues it can also create the social and privacy issues. Therefore it is a serious issue and has to find an effective solution using innovative technique. In this context the entire effort is made for improving the existing infrastructure of phishing attack recognition. In this paper the main aim is to explore and study the phishing attack and their key issues. In addition of that the aim is to find the recent valuable contributions in the domain of phishing attack detection and their prevention [2]. In this context a survey of recent literature and methodologies are conducted to understand and provide an effective solution for the phishing issues.

Background

This section provides the information about the different terminology that is used in this paper.

Phishing: phishing is a cyber crime; in this act the attacker is tries to still the confidential and sensitive information from a target user for example credit card details, bank account information. In this context attacker prepare a false web page or false login page and the link (URL) shared with the target user. If the target user is not aware about the online frauds then they pass their information and trapped in the phishing [3].

PhishTank database: phish tank is an authentic URL repository where the different organizations and institutions who are working for cyber security report the phishing cases by using some different information. Therefore the phish tank also offers the experimental real world data for experimentation and anti-phishing techniques design. The following key attributes are published by the phish tank agency [4].

1. **Phish ID:** that denotes the ID assigned by phish tank database for recognizing the phishing attack.
2. **URL:** that demonstrates the URL which is used to deploy the phishing attack.

Corresponding author:* **Rahul Patel
Shri Vaishnav Institute of Information Technology SVVV Indore

3. **Phish detail URL:** the details about the phishing URL is available online in this link.
4. **Submission date:** the date of reporting this URL as phishing is given in this attribute
5. **Verification time:** the time when it is verified as the phishing URL
6. **Online:** that shows the detection type of URL
7. **Target:** the target company or organization for deploying the attack

Data mining: the data mining is a technique which is used to evaluate and analyze the data. The analysis of data is aimed to identify the fruitful data relations, features and coefficients by which similar kinds of patterns can be identified. The data mining techniques can be supervised and unsupervised. In supervised techniques the pre-labeled data is used to train the target algorithms and then the algorithm can be used for recognition purpose. On the other hand in unsupervised learning the algorithms are directly applied on data for pattern recovery [5].

Malicious URL: the web URLs are used to target a resource in web. But the given URL not behaving in regular manner for targeting the required resource can be termed as the malicious URL. These URLs are developed using the multiple redirections, for downloading the malwares and other suspected techniques [6].

URL Classification: the URL (uniform resource allocator) is the address of web page resources in World Wide Web. The URLs are the unstructured form of data when it is collected in dataset for classification purpose. That data cannot be used directly for classification with an algorithm. Therefore the different heuristics or feature computation techniques are utilized for finding the URL properties. After that these URLs are become classifiable [7].

Association rule mining: the association rule mining is also known as frequent pattern mining technique. In this kind of algorithm the frequently occurred patterns are identified on a transactional database. That technique provides the definition, how two data items are associated with each other. There are a number of different kinds of association rule mining algorithms available i.e. apriori, FP-tree and others [8].

Classification: the classification algorithms are the supervised learning algorithms in data mining. That techniques are working in two major phases, in first phase pre-labeled data is used to train the algorithms. Using the pre-labeled data system prepares the data model that can be a mathematical model. After learning the developed data model is used to classify unlabeled data. That phase is known as testing or recognition or classification, during this the trained algorithm predict the class labels on which the algorithm trained [9].

Clustering: the clustering is an unsupervised learning approach, which is used to group the data. The clustering techniques usages the similarity or distance matrix to conclude the data instance clusters. That technique works as the filtering algorithm to group data elements efficiently. That technique works unsupervised manner thus the class labels are not required during the algorithm process [10].

Literature Review

This section provides the recent contributions and research efforts in the directions of designing effective anti-phishing system or phishing URL recognition system.

The proposed works give a transformative calculation based AI approach for pernicious URL characterization. ***Anshika Bansal [11]*** use BAT calculation for highlights extraction shapes all URLs from the URL list. It incorporates distinguishing and deciding the highlights based way to deal with arrange various kinds of malignant URLs. The proposed methodology utilized the extricated information to Support Vector Machine classifier for the order of testing information to names as favorable, spam and phishing.

A "noxious site page" alludes to a page that contains a malevolent substance that can misuse a customer side PC framework. A malignant site might be utilized as a weapon by cybercriminal to abuse different security dangers, for example, phishing, drive-by-download and spamming. Vindictive Web destinations are an obstacle in transit of Internet security. What's more, utilized as a weapon to mount different security risk like phishing, drive-by-download and spamming. To deal with there is a need to build up a programmed framework to a perceived noxious site. To determine identification models for pernicious website pages, recognizing highlights of considerate and noxious site pages are examined. ***Ghanshyam Sen et al [12]*** gives a way to deal with locate the malignant URL utilizing PSO approach.

Malevolent URLs are wide wont to mount various digital assaults together with spamming, phishing and malware. Discovery of noxious URLs and recognizable proof of risk assortments territory unit critical to frustrate these assaults. Knowing the kind of danger licenses estimation of the seriousness of the assault and receives a decent advance. Existing procedures more often than not see vindictive URLs of one assault kind. Amid this paper, ***R. Hamsa Veni et al [13]*** tend to propose approach utilizing AI to see malignant URLs of all the well-known assault assortments and build up the character of assault a pernicious location attempts to dispatch. The strategy utilizes a scope of discriminative alternatives together with issue properties, interface structures, website page substance, DNS data, and system traffic. A few of those choices are novel and amazingly viable.

Enormous online informal communities with a huge number of dynamic clients are progressively being utilized by Cyber hoodlums to spread vindictive programming (malware) to misuse vulnerabilities on the machines of clients for the individual increase. Twitter is especially powerless to such action as, with its 140 character limit, usually for individuals to incorporate URLs in their tweets to connection to increasingly point by point data, proof, news reports, etc. URLs are regularly abbreviated so the endpoint isn't clear before an individual taps the connection. Cybercriminals can abuse this to engender noxious URLs on Twitter, for which the endpoint is a malevolent server that performs undesirable activities on the individual's machine. This is known as a drive-by-download. In this paper, ***Pete Burnap et al [14]*** build up a machine grouping framework to recognize pernicious and considerate URLs close to the URL being clicked (for example 'continuous'). They train the classifier utilizing machine

movement logs made while communicating with URLs removed from Twitter information gathered amid a huge worldwide occasion – the Superbowl–and test it utilizing information from another enormous game – the Cricket World Cup. The outcomes demonstrate that machine movement logs produce accuracy exhibitions of up to 0.975 on preparing information from the principal occasion and 0.747 on test information from a second occasion. Besides, creators analyze the properties of the scholarly model to clarify the connection between machine movement and malevolent programming conduct and assemble an expectation to absorb information for the classifier to delineate that extremely little example of preparing information can be utilized with just a little burden to execution.

Malignant URLs are unsafe to each part of PC clients. Recognizing of the pernicious URL is significant. As of now, recognition of pernicious site pages procedures incorporate boycott and white-list technique and AI arrangement calculations are utilized. In any case, the boycott and white-list innovation are futile if a specific URL isn't in the rundown. In this paper, ***Rajesh Kumar et al [15]*** propose a multi-layer model for identifying vindictive URL. The channel can legitimately decide the URL via preparing the limit of each layer channel when it achieves the edge. Something else, the channel leaves the URL to the following layer. They likewise utilized a guide to confirm that the model can improve the precision of URL location.

Proposed Work

The data mining and its techniques are helpful in various applications for pattern recognition and classification. Different applications are usage the techniques of machine learning and data mining for decision making, prediction and classification in the domain of business intelligence, banking, stock market, health care and others. Therefore these techniques have the potential to work with critical situations where the decisions based on previous experience can be taken. The proposed work is intended to demonstrate the application of data mining in the domain of web/cyber security. In this context the proposed work is focused on exploring the phishing attacks and their causes.

The phishing attack is one of the crucial attacks which are deployed for innocent web users. Basically using different communication channels a malicious URL is delivered to the target user such by email, SMS or any other technique which looks genuine. The target user when visit the given URL and provide their information then attacker uses the given information for extracting the amounts from bank or credit card. Therefore URLs are the key tool of attacker by which he/she trap the innocent web users. In this context the proposed work is motivated to design a phishing URL classification technique using data mining algorithm. The proposed working model is demonstrated using the figure 1.

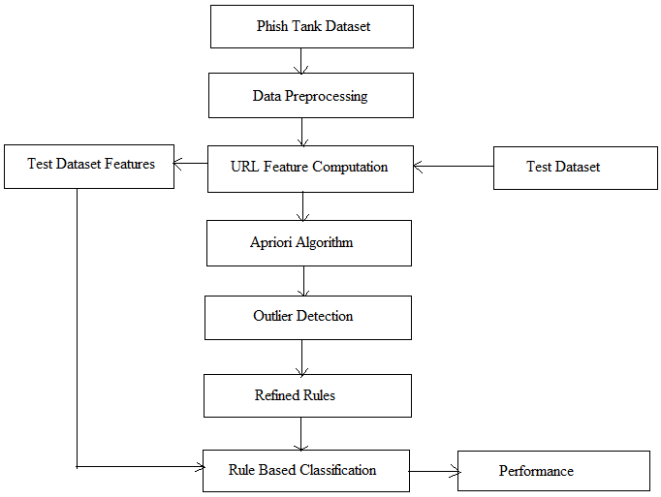

Figure 1 proposed system

An overview of the proposed system is given using figure 1. In this diagram the phish tank dataset is used for training of the system. The phish tank dataset contains 7 different attributes, among all the attributes are not much essential for proposed classification system design. Therefore in pre-processing phase the URL attribute remain preserved and other attributes are removed from dataset. The refined URLs from the phish tank database cannot be used directly with any classification or computational algorithm. Therefore the URLs are evaluated with the different heuristic functions to extract the features from the dataset. These features are explained in article [16].

After extraction of 14 different features from URLs the data is transformed into binary information using [16]. The encoded features are now used for developing the transactions which can be used with the association rule mining technique. Therefore apriori algorithm is used for extracting the features. The apriori algorithm helps to recover the association rules from the data. These rules are further pruned using the outlier detection technique. Finally the refined URLs are used for classification of the test URLs or unknown URLs. Before classifying the unknown URLs using the given association rule mining technique it is required to recover the features as extracted during the training phase. Finally the classified outcomes can be used for performance evaluation of the developed system.

CONCLUSION & FUTURE WORK

In this age of internet everyone becomes a consumer of internet services. Internet is a tool to offer services in users doorsteps. As the usages of internet increases a significant amount of online fraud and phishing is also increases. In this context a number of techniques are coming in existence but most of them are not much effective for preventing or accurately recognizing the phishing URL patterns. In addition of that some techniques are effective but a significant amount of computational overhead produces during classification. On the other hand there are some machine learning techniques are also available that are accurately recognizes the patterns of phishing URLs. Therefore the proposed work provides the study of data mining technique based phishing URL classification. The proposed technique is a promising system and can accurately classify the malicious URLs. In near future the proposed technique is

implemented using the relevant technology. Additionally to justify the effort placed their performance is also reported in near future.

References

1. Archana Saxena, Dr. R.D. Yadav, "IMPACT OF MOBILE TECHNOLOGY ON LIBRARIES: A DESCRIPTIVE STUDY", *International Journal of Digital Library Services,* Vol 3, Oct.–Dec. 2013, Issue—44
2. Long Cheng, Fang Liu and Danfeng (Daphne) Yao, "Enterprise data breach: causes, challenges, prevention, and future directions", Volume 7, September/October 2017 1 of 14 © 2017 The Authors. WIREs Data Mining and Knowledge Discovery published by John Wiley & Sons, Ltd.
3. AHMED ALEROUD, LINA ZHOU, "Phishing Environments, Techniques, and Countermeasures: A Survey", Computers & Security Volume 68, July 2017, Pages 160-196
4. Xiao Han, Nizar Kheir, Davide Balzarotti, "PhishEye: Live Monitoring of Sandboxed Phishing Kits", CCS'16, October 24 - 28, 2016, Vienna, Austria c 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM. ISBN 978-1-4503-4139-4/16/10
5. Ainhoa Goienetxea Uriarte, Enrique Ruiz Zúñiga, Matías Urenda Moris, Amos H. C. Ng, "How can decision makers be supported in the improvement of an emergency department? A simulation, optimization and data mining approach", Operations Research for Health Care 15 (2017) 102–122
6. Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma, "Detection of Malicious URLs using Machine Learning Techniques", *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, ISSN: 2278-3075, Volume-8 Issue-4S2 March, 2019
7. Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi, "Malicious URL Detection using Machine Learning: A Survey", arXiv:1701.07179v2 [cs.LG] 16 Mar 2017
8. Rahul B. Diwate , Amit Sahu, "Data Mining Techniques in Association Rule : A Review",) *International Journal of Computer Science and Information Technologies*, Vol. 5 (1) , 2014, 227 -229
9. Gaurab Tewary, "EFFECTIVE DATA MINING FOR PROPER MINING CLASSIFICATION USING NEURAL NETWORKS", *International Journal of Data Mining & Knowledge Management Process (IJDKP)* Vol.5, No.2, March 2015
10. Erich Schubert, Alexander Koos, Tobias Emrich, Andreas Zufle, Klaus Arthur Schmid, Arthur Zimek, "A Framework for Clustering Uncertain Data", Proceedings of the VLDB Endowment, Vol. 8, No. 12 Copyright 2015 VLDB Endowment 2150-8097/15/08.
11. Anshika Bansal, "Malicious Web URL Classification using Evolutionary Algorithm", *International Journal of Scientific Research & Engineering Trends,* Volume 3, Issue 5, Sept.-2017
12. Ghanshyam Sen, Himanshu Yadav, Anurag Jain, "An Approach to Detect Malicious URL through Selective Classification", *International Journal of Scientific Research & Engineering Trends*, Volume 3, Issue 4, July-2017
13. R. Hamsa Veni, A. Hariprasad Reddy, C. Kesavulu, "Identifying Malicious Web Links and Their Attack Types in Social Networks", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* © 2018 IJSRCSEIT | Volume 3 | Issue 4| ISSN: 2456-3307
14. Pete Burnap, Amir Javed, Omer F. Rana, Malik S. Awan, "Real-time Classification of Malicious URLs on Twitter using Machine Activity Data", 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM'15 August 25-28, 2015, Paris, France, ACM
15. Rajesh Kumar, Xiaosong Zhang, Hussain Ahmad Tariq, Riaz Ullah Khan, "Malicious URL Detection Using Multi-Layer Filtering Model", 978-1-5386-1010-7/17/$31.00 ©2017 IEEE
16. S. Carolin Jeeva and Elijah Blessing Rajsingh, "Intelligent phishing url detection using association rule mining", Hum. Cent. Comput. Inf. Sci. (2016) 6:10, DOI 10.1186/s13673-016-0064-3

How to cite this article:

Rahul Patel and Ananad Rajavat., 2019, Fight Against Phishing: A Data Mining Technique. *Int J Recent Sci Res.* 10(05), pp. 32494-32497. DOI: http://dx.doi.org/10.24327/ijrsr.2019.1005.3483

****